FACTION®

Limitless File Scale-Out in a Multi-Cloud World

By Dan Oustecky

08/05/2020

Contents

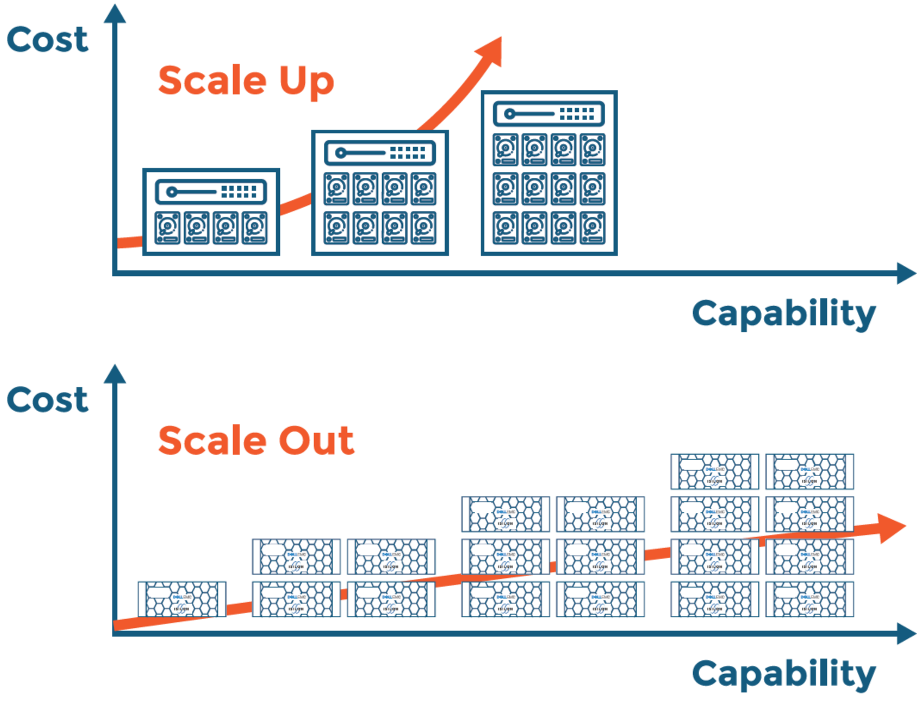
Big Data Requires Scalable Storage Systems

Big Data problems require Big Data-capable storage systems that can scale as your data grows. Traditional NAS arrays usually scale only one way: up. Although scaling up has worked well in the past, it's often ill-equipped for the demands of today's data. In contrast, scale-out architecture offers new advantages. In this paper, I'll review the limitations of "scale up" architecture and how "scale-out" architectures address those same issues.

When Scaling Up Breaks Down

Scaling up a storage array means adding additional capacity or performance to an existing controller or existing set of Highly Available (HA) controllers. To use an automotive analogy, you could add performance to your existing car by adding a cold air intake or a forced induction system. Meanwhile, if you wanted to add capacity to fit more stuff, you could add a rooftop box or a trailer. With data storage, the scaling-up process is the same. You can add more disk shelves to increase your capacity, and there may be limited ways, like increasing memory or adding cache, to boost performance, but traditional NAS arrays will always bump up against a ceiling.

Scale Up vs. Scale Out


Scale Up

Scale Out

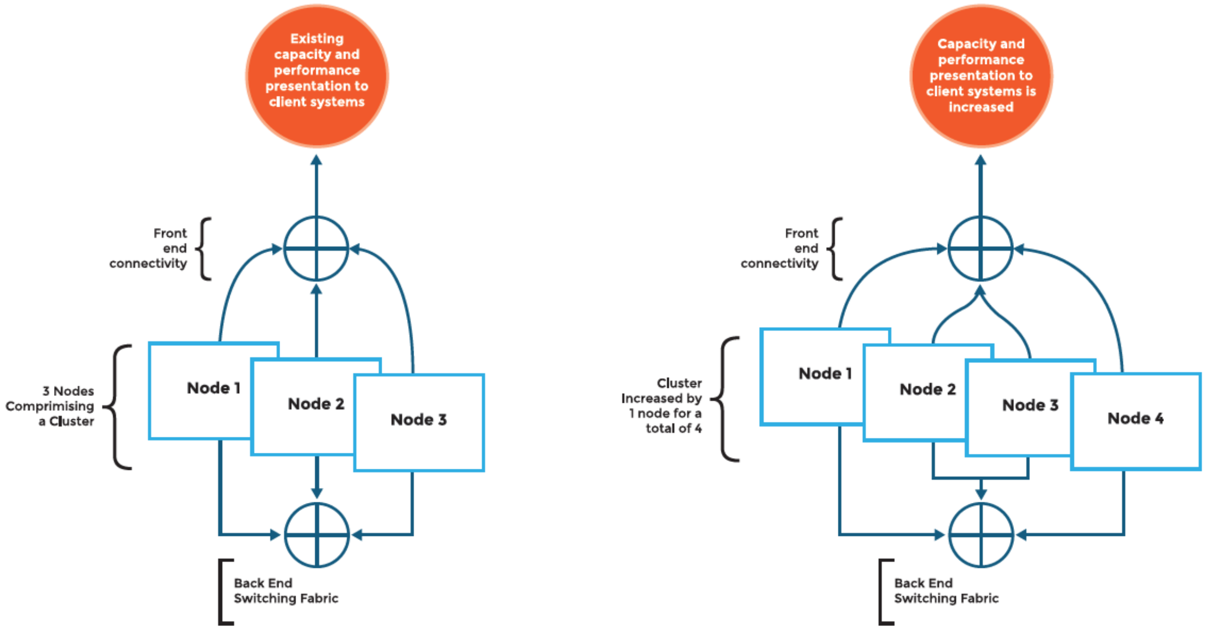
This limit usually happens when storage administrators and engineers can still add capacity, but there is effectively no way to increase performance. When controllers with the same performance level become responsible for more and more capacity, the performance per TiB ratio drops to the point where it may no longer be acceptable for the application and business requirements.

For example, if we have an array controller capable of 10,000 IOPS and 5GBps, doubling the storage capacity drops the performance per TiB in half:

Example Controller at Different Storage Capacities	IOPS	Throughput	Performance Per TiB
100 TiB	10,000	5GBps	50MBps/TiB 100 IOPS/TiB
200 TiB	10,000	5GBps	25 MBps/TiB 50 IOPS/TiB

The classic escape routes from this quandary require substantial compromises that demand increasingly more expensive controllers to support larger and larger capacities, driving the price per TiB ratios up. This approach may also create "array sprawl", where many independent arrays with siloed performance and capacity are deployed which creates pockets of capacity separated from your arrays with available performance. This design is fraught with inefficiency, operational complexity, and bloat. This model can have disastrous effects for workloads whose performance needs grow as the data set expands.

However, organizations can solve many of the problems of scale-up systems with a “scale-out” storage system that scales performance alongside capacity increases. Scale-out works by sharing a single file system across many nodes connected to a dedicated back-end (BE) switching fabric. As you add nodes with additional capacity, their compute and network capabilities are also added to the cluster, increasing aggregate performance capabilities under a single namespace.


Existing capacity and performance presentation to client systems
Front end connectivity
3 Nodes Comprimising a Cluster
Node 1
Node 2
Node 3
Back End Switching Fabric
Capacity and performance presentation to client systems is increased
Front end connectivity
Cluster Increased by 1 node for a total of 4
Node 1
Node 2
Node 3
Node 4
Back End Switching Fabric

Scale-out storage is suitable for data-intensive workloads across industries, including Advanced Driver Assistance Systems (ADAS), media and entertainment workflows, Scientific Computing, and IoT.

1.1 Where Does Scale-Out Storage Make Sense?

1.1.1 Autonomous Driving Demands a Scale-Out Model

Let's examine how Scale-Out is useful with a real-world example. Both IoT and ADAS generate continuous streams of sensor data that need to be processed. As more sensors come online, these data streams require additional storage and increasing performance to continue processing all of the data. The data processing stage is especially important to consider here as well. Scaling models for compute infrastructure were one of the first to transition to a cloud model, but getting your data to the cloud has not followed as quickly. Keeping data synced across clouds and with your primary data repository has become more challenging in traditional designs. This challenge can be compounded with siloed performance and capacity "islands" common with traditional NAS array deployments.

In a recent Blocks & Files interview, three autonomous vehicle (AV) industry aces estimated that daily accumulated storage capacity needs for a single consumer AV could range as high as 5 to 15 TiB. Autonomous taxis, which are being designed to operate as close to 24X7 as possible, are expected to require nearly 100TiB of daily data storage, and possibly up to 400TiB daily. The majority of that data will be processed and stored/buffered locally, but even experts are unsure of how much data will need to be phoned home for further processing or retention. For fleets of vehicles that number in the hundreds of thousands, even a small fraction of that capacity would overwhelm most enterprise arrays. For example, Tesla sold over 367,000 vehicles with this capability in 2019 alone. If only 100,000 of those vehicles have advanced driver aids enabled and in use, and only 0.1% of the 15TiB daily storage needs mentioned earlier is transferred, that could result in data uploads from vehicles totaling 1.6PiB per day!

1.1.2 Scaling for Media & Entertainment Demands

The performance needs for media and entertainment workflows, if not as daunting with capacity needs, is similarly intensive. With teams of engineers and artists simultaneously working on non-linear editing (NLE) client systems and massive rendering and transcoding jobs, the throughput needs can be astounding. Additional constant background load of media asset management services, meta-data tagging and updating, and a continual ingestion of raw creative material requires the distributed performance characteristics that only a scale-out storage platform can provide.

1.1.3 Other Use Cases

With IoT data being largely unstructured, scale-out storage systems make for a perfect match. Often object-based storage can be leveraged for such purposes, but meaningful performance can be difficult to achieve. Additionally, not all unstructured data is suitable for object access patterns or latency requirements. Scale-out NAS platforms can offer similar capacity and namespace scaling features of object storage but with a performance that far exceeds cloud provided capabilities especially with regard to latency. Additionally, if transform operations or object/file updates are part of your workflow, performance for these operations at scale is generally far superior with existing storage protocols not offered by object only services.

IoT data has additional challenges outside of simply collecting all of this data. IoT data can be pre-processed at the source or can be raw data from the source but the data is not being collected just to store. To derive value from this data, it needs to be processed and transformed into useful output or actionable inferences. Machine Learning and Artificial Intelligence (ML/AI) are used to sort through the data and pluck out useful insights. These ML/AI systems require huge amounts of data just to train. The return for these efforts is greater or more accurate results with larger data samples or sample sets.

Machine Learning and Artificial Intelligence (ML/AI) is another area where scale-out designs prove extremely effective. High concurrency rates from GPU farms can put tremendous pressure on the supporting storage systems. The distributed resources comprising a scale-out storage system are able to spread that load across all nodes. This ability often can extend beyond the IO operations themselves to include cluster-wide caching capabilities and distributed file locking allowing data scientists to train and utilize extremely large neural networks. ML/AI environments are often meticulously designed for performance and to prevent IO congestion. Scale out storage systems are often considered a prerequisite.

1.2 Technical Considerations

The term Scale-Out storage describes an architecture that scales capacity and performance with the addition of nodes, but there are a number of ways to accomplish that. Some vendors have enabled their traditional NAS arrays to work similarly to a fully distributed scale-out cluster by layering virtual file systems on top of existing RAID structures which can then be extended, manually, to additional controller pairs as they are added. While technically this is a scale-out design, you may find unexpected limitations in node count (more importantly the connectivity and processing capability of the nodes), or unexpected performance characteristics with your data, workload, or access pattern. Additionally, as a virtual file system overlaid on top of existing RAID structures, there are additional complexities to the initial configuration, scaling and ongoing maintenance.

A more precise description of Scale-Out architectures would augment the description above to include a fully distributed architecture. Many object stores meet this definition with only the access protocol and namespace being significantly different. Software data protection schemas obviate the need for traditional RAID structures and components. As a fully distributed architecture, there is no need to manually create space for your data, or create constituent or member volumes to feed your namespace. There is no concept of pool or aggregate ownership and no traditional RAID levels. Load balancing and data protection schemas are also far easier to digest, manage, and deploy.

Object Storage meets all the definitions of a scale-out storage architecture. Object storage generally differs not with the underlying hardware architecture but with the “file system” and method of access. Object storage platforms store files as logical “objects” that contain all of the data and metadata for that particular file. These objects have unique identifiers that allow for reference or retrieval without the need for a hierarchical file system tree producing a completely flat namespace. Object protocols (S3, Swift CDMI, etc.) are really HTTP RESTful API standards which can be at odds with access requirements. Object storage gateways may be offered from the object platform itself or can be deployed intermediary, as a translator of sorts, to allow for file or block protocol access,

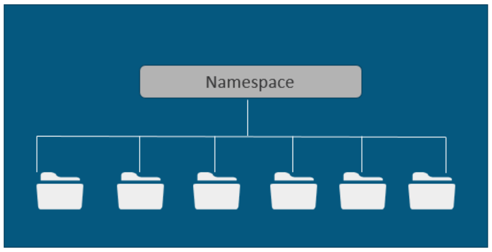


Figure 1 Flat namespace file system

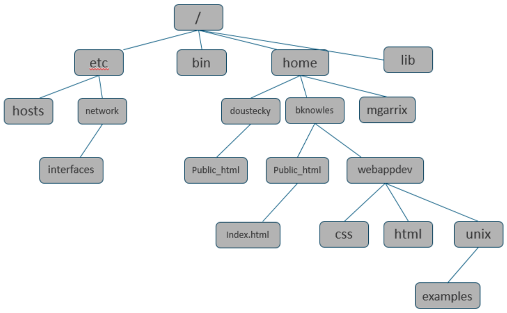


Figure 2 Hierarchal File System

but performance is a secondary consideration. The latency of the API requests will always hinder IOPS performance. This is likely a design to avoid if transactional writes are predominant. If alternative access protocols or methods are required for your workflow, scale-out NAS platforms offer access capabilities like S3 alongside NFS and SMB or for a more distributed storage access, HDFS can be leveraged.

Hadoop, using HDFS can especially benefit from the distributed architecture of a scale-out platform for tiered deployments. Big data analytics problems virtually require external storage systems that can act as both a name node and a data node. Hadoop clusters issue requests from job schedulers to worker nodes and data retrieval requests are load-balanced by a distributed name node service and pointed to the optimal data node for retrieval. This high integration provides an extra level of scale, efficiency, and performance to platforms and deployments that support it.

Let's consider a fully distributed scale-out design such as Dell EMC's PowerScale. PowerScale has a single, global namespace for the entire cluster. As you add nodes to the cluster, their capacity and performance capabilities are incorporated into the cluster. Data is protected from individual device or Field Replaceable Unit (FRU) failure not by legacy RAID structures but through software-based data protection. Data is striped across nodes and Forward Error Correction (FEC) stripes, based on Reed-Solomon provide parity to recalculate data from a failed node or disk. This functions very similar to RAID but is done in software and across all nodes in the cluster. Because of this, the protection schema does not need to exist at the physical hardware layer and is instead applied to the data or files, themselves. With the data protection schema acting on the data itself, you can protect your data in flexible and interesting ways.

Data protection being provided at the file system level does not preclude hardware failures. Node failures and disk failures need to be considered and expected. Software data protection and a distributed file system make this easy with selectable protection levels.

Protection Level	Description
+1n	Tolerate failure of 1 drive OR 1 node
+2d:1n	Tolerate failure of 2 drives OR 1 node
+2n	Tolerate failure of 2 drives OR 2 nodes
+3d:1n	Tolerate failure of 3 drives OR 1 node
+3d:1n1d	Tolerate failure of 3 drives OR 1 node AND 1 drive
+3n	Tolerate failure of 3 drives or 3 nodes
+4d:1n	Tolerate failure of 4 drives or 1 node
+4d:2n	Tolerate failure of 4 drives or 2 nodes
+4n	Tolerate failure of 4 nodes
2x to 8x	Mirrored over 2 to 8 nodes, depending on configuration

Data protection needs change over time. Because this protection is software-based, as protection requirements change you can realign protection levels to those data needs. Changes are applied at a per-file granularity level and without downtime or offline reconstruction. Additionally, as the industry moves forward and new protection levels become necessary, they are just a software update away.

Interestingly, this has some knock-on benefits. With a global namespace, flexible data protection schemas, and nodes that incorporate their capacity and performance into the cluster, maintaining spare disks is re-focused to maintaining spare capacity. With traditional RAID levels, when a disk fails, it either needs to be immediately replaced with a new disk or a hot spare needs to be allocated to rebuild the raid group. This is usually automatic with hot spares, but eventually the failed disk needs to be replaced to allow for the backfill of the consumed hot spare.

Distributed data protection incorporates all nodes and disks into the cluster, so there is no concept of physical spares. All resources are active. Instead, free or “slack” capacity within the pool provides the same function. At the detection of a failed disk, affected data, and only affected data, is immediately rebuilt by the Flex Protect job on alternate disks and nodes already in the cluster, consuming the free space equivalent of the data that was on the failed disk. This has an improved Time to Recovery (TTR) aspect of not requiring reconstruction of an entire RAID group, lowering the time necessary to regain full protection. Up to four virtual hot spares can still be configured if desired, reserving equivalent capacity on the cluster so that the total capacity of the cluster is not reduced while a disk is failed. Additionally, IO stress from rebuilding data or parity is distributed across the entirety of the cluster. Reconstruction hotspots, which can degrade front end performance, are not possible. Instead, there is virtually no application impact to disk failures and recovery time is significantly lowered.

An additional benefit comes in the form of raw capacity utilization. The difference between the raw capacity and usable capacity is controlled by the data protection methodology used. In a distributed scale-out cluster, data protection is performed in software and parity is applied to the individual files and not to RAID groups comprising a pool. All disks are used for both parity and data but interleaved based on how files are distributed across nodes and disks. Raw storage utilization is higher than many traditional NAS architectures. This introduces the concept of capacity flexibility, which is available with policy-based data protection levels applied to specific files or file sets.

Scaling capacity and performance can be significantly easier to accomplish than one might expect. Jobs like Autobalance, ensure that all nodes have the same capacity utilization. SmartFail continuously monitors for Error Correction Codes (ECC) and SMART failures, proactively removing a failing disk, redistributing data across the cluster without reconstructing. Both features rely on flexible data protection capabilities to provide a near fully automated ongoing care and feeding.

Upgrades and performance/capacity augmentations are similarly made easier. Depending on the configuration, nodes can be automatically added to the cluster and capacity allocated to appropriate pools. Software upgrades are also simple to accomplish. Various upgrade methods allow for tailoring maintenance to your needs. You can run a “simultaneous” upgrade, which upgrades and restarts all nodes in the cluster at the same time. This has an obvious impact on data availability and requires serious consideration. A “rolling” upgrade or newly added “parallel” upgrade, is likely a more appropriate choice for production. Rolling upgrades install new code and reboot nodes one at a time, waiting for each node to complete before beginning the next. Parallel upgrades are intended for larger clusters where individual nodes, grouped into neighborhoods, can be simultaneously upgraded with nodes from other neighborhoods, never impacting cluster data protection more than a rolling upgrade would. This can be a huge time saver for larger clusters versus the rolling method, where a cluster upgrade can last an entire workday or overnight maintenance window.

Scale-out storage architectures are not perfect and some workloads may not be appropriate. While the transactional performance of scale-out NAS clusters is very likely to exceed cloud object storage, a high-performance traditional array is likely to surpass a scale-out cluster under the same circumstances. These high Input Output Per Second (IOPS) workloads like high-frequency trading, enterprise databases, or order processing systems may see higher than expected latencies because data has to be federated across the nodes in the cluster. Additionally, High IOPS are often correlated with lower IO size, which, while faster to process, can often be hamstrung by the higher latency required to spread the data evenly across all nodes. While this latency is small, it exists and can become the limiting factor when pushing the upper limits of small IO performance or maximum IOPS. Mitigations exist, such as allocating a dedicated pool for intensive workloads allowing IOPS focused workloads to push nodes and disks to their full transactional potential but this highly specialized type of load should often be approached with an alternate storage architecture.

The Opportunity and the Path to Get There

With a sufficiently performant scale-out cluster backing your big data workload, the performance question often turns back to the cloud. In order to realize the potential provided, sufficiently scalable compute infrastructure is needed to process growing datasets. Having the latest compute resources and most advanced data processing services available at the flip of a switch (more aptly the swipe of a credit card) is irresistibly appealing. The opportunity to leverage your data requirements for a competitive edge by enabling cloud arbitrage is an exciting thought. Normally considered a risk or liability, the requirement to process your data can allow for an exciting environment where you can cherry-pick the best services, solutions, and resources from any cloud. The key to success with this architecture is low latency achieved through cloud adjacency.

Storage network topologies for a hybrid or multi-cloud access design are trade-offs. Choosing to access remote, on-premises storage from a hyper-scale cloud still presents challenges with regard to latency, but some of that performance can be made back up through brute force. Latency has a dramatic effect on per-flow or per-stream performance/throughput but aggregating many flows/streams together can still provide for high overall throughput scenarios. This is not fixing the performance issues rooted in high latency but allows for leveraging concurrency or threading or queue depth. Scale-out platforms are especially adept at processing and fulfilling this level of concurrency.

This is easy to comprehend when looking at the scaling architectures of scale-out NAS versus traditional NAS arrays. In a scale-out design, as capacity increases so does the node count and often front end (FE) connectivity. Traditional arrays are generally monolithic and may only increase node counts after previously scaling capacity numerous times. At best this leads to a staircase function type of asymmetric scaling which is significantly harder to maintain and balance as performance needs rise. Scale-out architectures allow for a much more linear performance and capacity growth model. Unfortunately, transactional or single data flow workloads will always suffer when the network connectivity is highly latent, regardless of the storage architecture.

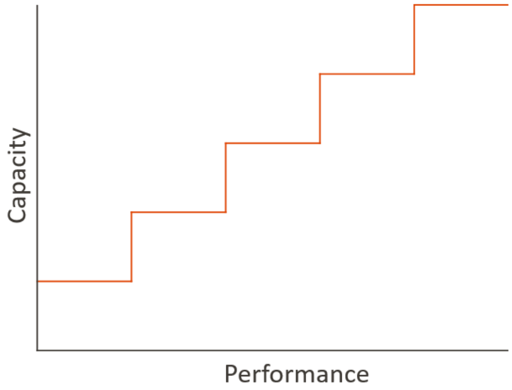


Figure 3 Staircase functions have big steps to next level of performance and capacity

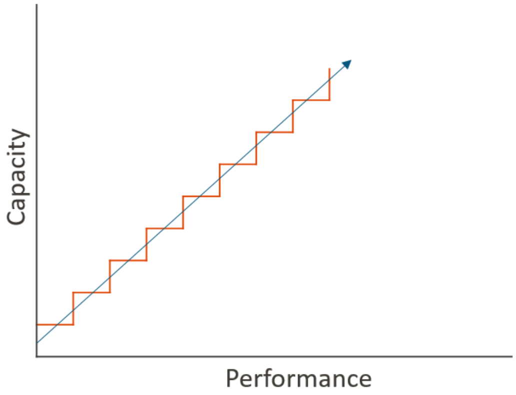


Figure 4 File Scale-Out achieves a more linear distribution

Scale-Out NAS meets Multi-Cloud

The best solution for latent connectivity to a cloud is adjacency. This concept becomes more important as more clouds or cloud services that require data access are brought online. A simple, and often adequate, solution is migrating data into a cloud provider. This generally closes the door on performant, affordable multi-cloud access. Superior resources or services from alternate cloud providers are walled off from your data, application, or workforce. Many of those services can dramatically increase efficiencies or decrease cost when compared to on-premises solutions. Amazon Kinesis or Azure Stream Analytics are great examples and these services are best solutioned with low latency data access that remote data lakes can't provide. This could mean leveraging ADLS or S3 and forklifting data to the cloud. Potentially duplicating data, but definitely diminishing flexibility, utilization of capacity, and performance to other clouds. The same dataset may not be usable across clouds or by on-premises resources and even when possible, cost, performance, or the method of access can be incongruent with your workflow.

Co-locating a scale-out NAS or contracting platform services at the cloud edge is the preferred adjacency solution, offering performant access from any cloud provider as well as simple DR and backup solutions either back to your on-premises environments or to a tertiary provider. The same cloud services afforded by forklifting your data into the cloud can easily be made available by placing your data at the edge of the cloud instead of in it. The advantage is that your dataset can be accessed from any cloud and multiple clouds can access and process the same data set in real-time becoming a fully functional multi-cloud data access solution. With a scale-out storage platform providing scalable performance and capacity, you can easily grow your data lake. New cloud compute resources to process your data are just as simple to deploy and enabling access from alternate clouds can be as easy as requesting access to that cloud's network onramps and then connecting the network edge of your new cloud to the virtual links provided.

Conclusion

Whether accessing locally on-premise or across a multi-cloud environment, a scale-out storage architecture based on fully distributed configuration can offer immense scale and performance for numerous applications and market verticals. Massive ADAS and IoT workflows requiring high aggregated throughput ingestion capabilities can scale wider and more linearly on a scale-out platform than any traditional NAS can provide. Likewise, media and entertainment workflows can leverage both on-box performance afforded with a distributed design as well as a multitude of access patterns.

Scale out NAS storage platforms afford numerous advantages over traditional NAS deployments for the right workloads. These features are all derivative functions of the distributed design and flexible data protection schemas. Improved capacity utilization, operational efficiency, and ease of management as capacity and performance scale are all important examples. While object platforms are often looked to for high scale needs, performance is not likely to be a primary concern. Access patterns and latency requirements will likely be a determining factor when selecting a scale-out storage platform.

The exceptional pairing of scale-out NAS solutions with cloud provided resources and services is a performance and efficiency powerhouse leveraging the best resources from the lowest latency possible in a multi-cloud capable data lake. This sort of configuration is fraught with pitfalls if attempted from an on-premises data source. New circuits, LOAs for cross connects and drops, delayed telco implementation, and, of course, high latency can all be avoided by moving your data source to the cloud edge. Likewise, when compared with forklifting your data into a cloud, deploying a multi cloud enabled strategy is far easier with scale out storage at the cloud edge and more flexible and performant across cloud providers than native cloud storage solutions.

About Faction

Faction is a leading multi-cloud data services provider that pioneered cloud-adjacent storage powered by patented technology that provides data access over low latency, high throughput connections to all the major clouds, including AWS, Azure and Google Cloud Platform. Faction is also a leading managed service provider for VMware Cloud on AWS, including disaster recovery and production deployments, and was the first to offer cloud-attached storage solutions that integrate natively with VMware Cloud on AWS. Faction's private and multi-cloud platforms give clients the ability to move, access, scale and protect data between clouds, without the fear of cloud lock-in. Faction, a VMware Cloud Verified provider, is also recognized as an Advanced AWS Consulting Partner and VMware Premier Cloud Provider. Follow Faction on Twitter (@FactionInc), LinkedIn, and BrightTALK. For more information, please visit www.factioninc.com.

SPEAK WITH A
CLOUD SPECIALIST

☎ +1 855 532 4734

✉ getcloud@factioninc.com